PERGAMON Computers and Mathematics with Applications 39 (2000) 43–48

An International Journal
computers & mathematics
with applications

www.elsevier.nl/locate/camwa

Evaluation of Some DNA Cloning Strategies

B. TANG*
Department of Mathematics
Arizona State University
Tempe, AZ 85287, U.S.A.

Abstract—By considering a DNA molecule as a random sequence of four letters, a mathematical model of two important steps in DNA cloning processes is described: degradation of large DNA molecules using restriction enzymes, and replication of the resulting DNA fragments by viral vectors. The model is used to evaluate the feasibility of obtaining target gene loci using a large number of restriction enzymes, and to estimate the fraction of genomic DNA that is clonable.

1. INTRODUCTION

In this paper, the efficiency and feasibility of two strategies in DNA cloning are evaluated with simple mathematical models.

All DNA cloning processes involve degradation of large DNA molecules into smaller fragments which are inserted into viral carriers called vectors for replication. Thus, the actual cloning is performed on the DNA fragments. There are many known and unknown factors which render the cloning process unsuccessful. One of the known factors is size limitation on DNA fragments which can be inserted into the cloning vectors.

Each strand of the DNA double helix is a linear chain of nucleotides linked by chemical bonds. Mathematically it can be taken as a random sequence of four letters G, C, A, and T, representing the four nucleotides or bases: guanine, cytosine, adenine, and thymine. Degradation of DNA molecules into small fragments is commonly carried out with the help of *restriction enzymes* which cleave chemical bonds at specific sites called *restriction sites*. Restriction sites are short sequences of bases, often four or six. For example, the sequence $G-A-A-T-T-C$ is the restriction site for the restriction enzyme Eco $R1$ which cleaves the bond between G and A in that sequence. The degradation process using a particular restriction enzyme is called a *complete digest* if all restriction sites are cleaved, and a *partial digest* if only a portion of the restriction sites are cleaved. (See, for example, [1] for a more detailed description of the cloning process.)

The location of the restriction sites obviously determines the size of the fragments after a digest. Experimental results reported by Hamer and Thomas [2] indicate that the distribution of restriction sites can be assumed to be random. Based on this randomness assumption, and taking into account of size limitation by the cloning vectors as the only cause of failure of the cloning process, simple probabilistic and combinatorial models can be constructed to evaluate the feasibility of cloning genes using complete digests with different enzymes [3], and to estimate the fraction of a genome that is clonable by complete as well as partial digest [4,5].

*Deceased.

0898-1221/00/$ - see front matter Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX
PII: S0898-1221(00)00106-1

This paper is organized as follows. Mathematical descriptions and notations of the DNA molecule and the cloning process are given in Section 2. Cloning genes by complete digestion is studied in Section 3, and the fraction of a genome that is unclonable is estimated in Section 4.

2. DNA CLONING—A MATHEMATICAL DESCRIPTION

The number of bases in a DNA fragment is referred to as the *length* of that fragment, and the units of measurement are bp and kbp—one base pair and one thousand base pairs. (A base is often called a base pair because of the pairing of $G-T$ and $A-T$ in the double helix.) The minimum and maximum length of DNA fragments that can be inserted in a particular vector are denoted by, respectively, $L+1$ and $L+R$. For bacteriophage λ vectors, which are widely used for cloning, $L \cong 9\,\text{kbp}$ and $R \cong 14\,\text{kbp}$.

The mean frequency of occurrence of restriction sites of a particular restriction enzyme, denoted by p, can be estimated by letting $p = 1/\mu$, where μ is the mean fragment length after complete digest with that restriction enzyme. The value of μ can be determined experimentally, and there are various theoretical estimates of μ based on Markov chain analysis on the sequence in the restriction site (see, for example, [5–8]). For many commercially available restriction enzymes, $\mu \sim 10^3$ to 10^4, so there is insignificant error in assuming that every bond between two bases has equal probability p of being cleaved by the restriction enzyme, even though strictly speaking two bonds which can be cleaved must be at least some distance apart (6 bp in the case of Eco $R1$). We will also loosely say that every bond has probability p of being a restriction site.

In order to describe the location of restriction sites, the bonds between the bases are indexed in the following way. Fix a particular portion S of the DNA molecule, which can be a gene or one single base. The r^{th} bond to the left of the first base on the left (5' side) of S is labeled $-r$, and the s^{th} bond to the right (3' side) of the first base is labeled s. The location of the i^{th} restriction site on the 5' side of S is x_i, and the location of the j^{th} restriction site on the 3' side of S is y_j. The set $(\mathbf{x},\mathbf{y})_S = \{\dots, x_2, x_1, y_1, y_2, \dots\}$, where $\cdots < x_2 < x_1 < 0 < y_1 < y_2 < \cdots$, gives the location of all the restriction sites on the DNA strand with respect to the first base of S, and is called the *rsc—restriction site configuration* (with respect to S). When no confusion arises, the subscript S is dropped.

The probability that a DNA molecule has a particular rsc $(\mathbf{x},\mathbf{y})$ (with respect to S) can be computed easily. For example, the probability that the rsc is such that $x_1 = A$ and $y_1 = B$, $A \le -1$ and $B \ge 1$, is

$$p^2(1-p)^{B-A-2},$$

since the $B-A-2$ bonds between x_1 and y_1 are not restriction sites.

A fragment obtained by cleaving the bonds, not necessarily restriction sites, at r and s, $r < s$, according to a particular rsc, is denoted by $[r,s]$. The length of the fragment, $|[r,s]|$, can be easily computed:

$$|[r,s]| = \begin{cases} s-r, & r < s < 0 \text{ or } 0 < r < s, \\ s-r-1, & r < 0 < s. \end{cases}$$

The fragment $[r,s]$ is *clonable* if $L+1 \le L+R$.

In a sample of DNA molecules prepared for digestion, every distinct molecule is present in a large number of identical copies ($\sim 10^9$). In a complete digest, all copies of the same DNA molecule are cleaved in exactly the same fashion. According to a particular rsc $(\mathbf{x},\mathbf{y})$ of the DNA molecule, the fragments obtained must be of one of the following forms: $[x_{i+1}, x_i]$, $[x_1, y_1]$, $[y_j, y_{j+1}]$, $i \ge 1$, and $j \ge 1$. In a partial digest, it is possible to obtain the fragment $[a,b]$ so long as $a < b$ and the bonds located at a and b are restriction sites. If there is a restriction site at c, $a < c < b$, production of the fragment $[a,b]$ does not eliminate the possibility of obtaining a fragment $[c,d]$ or $[e,c]$ where e and d are restriction sites, $e < a$ and $b < d$, since the action of the restriction enzyme can yield $[c,d]$ or $[d,e]$ from other copies of the same DNA molecule in the sample.

3. CLONING TARGET GENES BY COMPLETE DIGEST

The following simple strategy utilizing completely digested DNA is suggested for cloning genetic loci of different sites [3]. DNA fragments are generated in a number of digests, each of which employs different restriction enzymes, and are then inserted in bacteriophage λ cloning vectors. The feasibility of the strategy in producing clonable fragments containing the locus of interest can be estimated mathematically.

Let n be the number of restriction enzymes employed, and G the length of the gene locus of interest. For successful cloning it is obvious that G must satisfy $G \leq L + R$. Any bond between two bases in the DNA molecule has probability p_i of being a restriction site of the i^{th} enzyme. It is assumed that the gene locus is "centrally" located so it is equally likely for the gene to be the 5' end or the 3' end of a fragment after digestion.

Denote by $\mathbf{P}_i$ the probability that the gene locus of length G is contained in a clonable fragment produced by the i^{th} restriction enzyme. $\mathbf{P}$, the probability that the gene locus is contained in a clonable fragment produced by any one of the n restriction enzymes, is given by

$$\mathbf{P} = 1 - \prod_{i=1}^{n} (1 - \mathbf{P}_i).$$

Estimation of $\mathbf{P}_i$ follows directly from the theoretical study of depolymerization by Kuhn [9], Montroll and Simha [10], etc. Since all restriction sites are cleaved in a complete digest, the gene locus is contained in a fragment of length ℓ after complete digest with the i^{th} enzyme if and only if $G \leq \ell$ and the rsc $(\mathbf{x}, \mathbf{y})_G$ of the enzyme satisfies

$$y_1 = x_1 + 1 + \ell, \qquad -1 \geq x_1 \geq G - \ell - 1.$$

The probability that the locus is contained in a fragment of length ℓ is the probability that the DNA molecule has a rsc satisfying the above properties

$$\sum_{x_1=-1}^{G-\ell-1} p_i^2 (1 - p_i)^{\ell-1} = (\ell - G + 1) p_i^2 (1 - p_i)^{\ell-1}.$$

Fragments of length ℓ containing the locus are clonable if and only if $L + 1 \leq \ell \leq L + R$ when $G < L + 1$, and if and only if $G \leq \ell \leq L + R$ when $G \geq L + 1$. Therefore,

$$\mathbf{P}_i = \begin{cases} p_i^2 \displaystyle\sum_{\ell=1}^{L+R} (\ell - G + 1)(1 - p_i)^{\ell-1} p, & G < L, \\ p_i^2 \displaystyle\sum_{\ell=L}^{L+R} (\ell - G + 1)(1 - p_i)^{\ell-1}, & L \leq G \leq L + R. \end{cases}$$

For small p_i and large L and R, $1 - p_i \cong \exp(-p_i)$, $L \cong L + 1$, and the summations can be approximated by integrals

$$\begin{aligned} \mathbf{P}_i &\cong \begin{cases} p_i^2 \displaystyle\int_L^{L+R} (\ell - G + 1) \exp(-p_i(\ell - 1))\, d\ell, & G < L, \\ p_i^2 \displaystyle\int_G^{L+R} (\ell - G + 1) \exp(-p_i(\ell - 1))\, d\ell, & L \leq G \leq L + R, \end{cases} \\ &= (\alpha p_i + 1) \exp(-p_i(G + \alpha)) - ((L + R - G)p_i + 1) \exp(-p_i(L + R)), \end{aligned} \tag{3.1}$$

where $\alpha = \max(0, L - G)$.

Direct computation shows that $\frac{d\mathbf{P}_i}{dG} < 0$, $1 \leq i \leq n$, hence the feasibility of cloning a gene locus decreases as its size increases. On the other hand, the larger number of restriction enzymes used the more likely a gene locus can be cloned. Obviously as G gets close to the upper size limit of clonable fragments, the probability of the gene being cloned will remain very low even if many restriction enzymes are used.

4. CLONING GENOMIC DNA BY PARTIAL DIGEST

Genetic information contained in unclonable fragments is lost in the cloning process. Partial digestion, even though technically more difficult, is much better than complete digestion in mapping and sequencing of genomes because a larger portion of the genome is contained in clonable fragments. We first estimate the fraction of a genome that is clonable, i.e., the fraction of bases that are contained in clonable fragments, after complete digestion using one restriction enzyme with mean fragment length p^{-1}, and then estimate the improvement using partial digestion. Our approach is different from the work of Seed *et al.* [11,12] who estimated the unclonable fraction by conditional probabilities.

All the genetic information of an organism can be assumed to be contained in one single large DNA molecule of size N. For *E. coli*, $N \cong 4.7 \times 10^6$, and for humans $N \cong 3 \times 10^9$. The fraction of the genome that is clonable is the same as the probability that an arbitrary base is contained in a clonable fragment after the digest, which in turn is the probability that the DNA molecule has a rsc $(\mathbf{x}, \mathbf{y})$, with respect to the arbitrary base, such that $L+1 \le |[x_1, y_1]| \le L+R$. This can be easily obtained by substituting G with 1 in (3.1) and subsequent approximation of $L-1$ by L for $L \gg 1$

$$\begin{aligned}&\text{fraction of clonable bases in a genome after complete digestion}\\ &\cong (Lp+1)\exp(-pL) - ((L+R)p+1)\exp(-p(L+R)).\end{aligned}$$

In a partial digest, a base, with corresponding rsc $(\mathbf{x}, \mathbf{y})$ for which $|(x_1, y_1)| \le L$, can still be contained in a clonable fragment so long as there exist $x_i < 0$ and $y_j > 0$, $i \ge 1$, $j \ge 1$, such that $L+1 \le |[x_i, y_j]| \le L+R$, and the restriction sites $x_{i-1}, \ldots, x_1$, $y_1, \ldots, y_{j-1}$ are not cleaved during digestion. The improvement over complete digestion can be evaluated quantitatively in the following way.

A base will never be contained in a clonable fragment after any partial digest if the corresponding rsc $(\mathbf{x}, \mathbf{y})$ is either one of the following two types:

$E1$: $|[x_1, y_1]| > L+R$,

$E2$: there exist k and m, $k \ge 1$ and $m \ge 1$, such that $|[x_k, y_m]| < L+1$, and if $i \ge 1$ and $j \ge 1$, $i \ne m$, $n \ne k$, either $|[x_i, y_j]| < L+1$ or $|[x_i, y_j]| > L+R$.

Let $P(E1)$ ($P(E2)$) be the probability that $(\mathbf{x}, \mathbf{y})$ is of type $E1$ ($E2$). Then $P(E1) + P(E2)$ is the probability that the base cannot be contained in a clonable fragment after any partial digest. Since there is no control over which restriction sites are cleaved in any one partial digest, a fragment $[a, b]$ may not be produced even though a and b are restriction sites. Therefore, $P(E1) + P(E2)$ is really only the lower bound of the fraction of unclonable bases in a genome after any one partial digest. In practice, with the large number of identical DNA molecules in a sample prepared for digestion, it can be assumed that the fragment $[a, b]$ will be produced for any pair of restriction sites a and b (see [4, Section 7]). Thus, $P(E1) + P(E2)$ is a reasonably good approximation of the fraction of the genome that is unclonable after a partial digest.

From the last section, the probability that an arbitrary base has a corresponding rsc $(\mathbf{x}, \mathbf{y})$ such that $|[x_1, y_1]| \le L+R$ is approximately

$$p^2 \int_0^{L+R} \ell \exp(-p(\ell-1))\, d\ell = 1 - (p(L+R)+1)\exp(-p(L+R)),$$

hence, $P(E1) = (p(L+R)+1)\exp(-p(L+R))$.

$P(E2)$ can be estimated when $L \le R+1$, and a lower bound can be obtained when $L > R+1$.

To estimate $P(E2)$ for $L \le R+1$, we first partition the set of all rsc of the form $E2$. Let A_t, $1 \le t \le L$, be the set of rsc of the form $E2$ for which there exist k and m, $k \ge 1$ and $m \ge 1$, such that $|[x_m, y_k]| = t$, and if $i \ge 1$ and $j \ge 1$, $i \ne k$ and $j \ne m$, either $|[x_i, y_j]| \le t$ or $|[x_i, y_j]| > L+R$.

Denote by $P(A_t)$ the probability that the rsc $(\mathbf{x}, \mathbf{y})$, with respect to an arbitrary base, belongs to the set A_t. Then

$$P(E_2) = \sum_{t=1}^{L} P(A_t).$$

LEMMA 4.1. *If $L \le R+1$ and $(\mathbf{x}, \mathbf{y}) \in A_t$, $1 \le t \le L$, there exists a unique pair (m, k), $m \ge 1$ and $k \ge 1$, such that $[x_m, y_k] = t$.*

PROOF. First note that $-t \le x_m \le -1$. Suppose (u, v) is another pair, $u \ge 1$ and $v \ge 1$, such that $|[x_u, y_v]| = t$ also. Assume without loss of generality that $x_m < x_u$. Then either $|[x_m, y_v]| \le t$ or $|[x_m, y_v]| > L+R$. However, since $y_v > y_k$, $|[x_m, y_v]| > t$, so it must be the case that $|[x_m, y_v]| > L + R$. But that implies $|[x_m, x_u]| > L + R - t \ge 2L - t - 1 \ge t - 1$, a contradiction.

LEMMA 4.2. *If $L \le R+1$ and $(\mathbf{x}, \mathbf{y}) \in A_t$, $1 \le t \le L$, with $[x_m, y_k] = t$, $x_{m+1} < y_1 - (L+R) - 1$, and $y_{k+1} > x_1 + (L + R) + 1$.*

PROOF. Note that $|[y_1, y_k]| \le t-1$, and either $|[x_{m+1}, y_1]| < t$ or $|[x_{m+1}, y_1]| > L+R$. If $x_{m+1} \ge y_1 - (L+R) - 1$, $|[x_{m+1}, y_1]| \le L+R$ and so it must be the case that $|[x_{m+1}, y_1]| < t$. On the other hand $|[x_{m+1}, y_k]| > L+R$, so $|[y_1, y_k]| = y_k - y_1 = |[x_{m+1}, y_k]| - |[x_{m+1}, y_k]| > L+R-t > R > t$, a contradiction. Similarly, it can be shown that $|[x_m, x_1]| > t$ if $y_{k+1} \le x_1 + (L + R) + 1$.

COROLLARY 4.3. *Suppose $L \le R + 1$ and $(\mathbf{x}, \mathbf{y}) \in A_t$, $1 \le t \le L$, with $|[x_m, y_k]| = t$. If $x_m < a < x_1$, $|[a, y_j]| < t$ for $1 \le j \le k$, and $|[a, y]| > L + R$ for $j > k$, and if $y_1 < b < y_k$, $|[x_i, b]| < t$ for $m \le i \le 1$ and $|[x_i, b]| > L + R$ for $i > m$.*

Consequently, $(\mathbf{x}, \mathbf{y}) \in A_t$ if and only if there exist a pair (m, k), $m \ge 1$ and $k \ge 1$, such that the bonds located at x_m, x_1, y_1, y_k, where $-t \le x_m \le -1$, $x_m + 1 \le x_1 \le -1$, $1 \le y_1 \le y_k - 1$, $y_k = x_m + t + 1$, are restriction sites, and the bonds located at a and b, where $x_1 + 1 \le a \le -1$, or $y_1 - (L + R) - 1 \le a \le x_m - 1$, and $1 \le b \le y_k - 1$, or $y_k + 1 \le b \le x_1 + (L + R) + 1$, must not be restriction sites.

THEOREM 4.4. *If $L \le R + 1$, $P(A_t) = (1-p)^{2(L+R)-t-1} p^2 g(t)$, where $g(t) = p^2 t^3/6 + p(1 - p/2)t^2 + (p^2/3 - p + 1)t$.*

PROOF. If $(\mathbf{x}, \mathbf{y}) \in A_t$, there are $2(L + R) - t - 1$ bonds which must not be restriction sites. Therefore,

$$P(A_t) = \sum_{x_m=-t}^{-1} \sum_{x_1=x_m}^{-1} \sum_{y_1=1}^{y_k} (1-p)^{2(L+R)-t-1} p^{\eta},$$

where $y_k = x_m + t + 1$, and

$$\eta = \eta(x_m, x_1, y_1) = \begin{cases} 2, & x_1 = x_m \text{ and } y_1 = y_{k'}, \\ 3, & x_1 = x_m \text{ and } y_1 < y_{k'}, \text{ or } x_1 > x_m \text{ and } y_1 = y_{k'}, \\ 4, & x_1 > x_m \text{ and } y_1 < y_k. \end{cases}$$

Direct counting gives

$$\begin{aligned} P(A_t) &= (1-p)^{2(L+R)-t-1} p^2 \sum_{x_m=-t}^{-1} \left[1 + (t - 1_p - (x_m + 1)(x_m + t) p^2\right] \\ &= (1-p)^{2(L+R)-t-1} p^2 \left(\frac{p^2 t^3}{6} + p\left(1 - \frac{p}{2}\right) t^2 + \left(\frac{p^2}{3} - p + 1\right) t \right). \end{aligned}$$

$P(E2)$ can then be estimated as simply

$$\begin{aligned} P(E_2) &= (1-p)^{2(L+R)-1} p^2 \sum_{t=1}^{L} (1-p)^{-t} g(t) \\ &\cong (1-p)^{2(L+R)} p^2 \int_0^L \exp(pt) g(t)\, dt. \end{aligned}$$

Upon integration we obtain

$$P(E2) = F(L,R,p) \cong \frac{1}{6}p^2L^2(pL+3)\exp(-p(L+2R)).$$

Hence, if $L \leq R+1$,

$$\text{fraction of clonable bases in a genome after partial digestion}$$
$$\cong 1-(p(L+R)+1)\exp(-p(L+R)) - \frac{1}{6}p^2L^2(pL+3)\exp(-p(L+2R)). \tag{4.1}$$

If $L > R+1$, the above description of rsc's in A_t, given in Lemma 4.2, is no longer valid since it is not necessary that $x_{m-1} < y_1 - (L+R) - 1$, and $y_{k+1} > x_1 + (L+R) + 1$. Therefore,

$$P(A_t) \geq (1-p)^{2(L+R)-t-1}p^2g(t)$$

and $P(E2) \geq F(L,R,p)$. Thus, $1-(p(L+R)+1)\exp(-p(L+R)) - F(L,R,p)$ is an upper bound for the fraction of clonable bases in the genome.

Monte-Carlo simulations indicate that (4.1) is a very good estimate of the fraction of clonable bases in a genome, even for $L > R+1$ (see [4]). As a comparison, consider complete and partial digest with the restriction enzyme Eco $R1$, for which $p \cong 1/3000$, and using bacteriophage λ vectors ($L \cong 9\,\text{kbp}$, $R \cong 14\,\text{kbp}$) for cloning. With complete digest, the fraction of clonable bases in a genome is approximately 19.5%, and with partial digest, the fraction is drastically increased to 99.6%. Naturally, the fraction of a genome that is clonable will be reduced when factors in the cloning process other than size limitation of the cloning vectors are taken into account, but for the purpose of retaining as much genetic information as possible, the advantage of partial digestion over complete digestion has been clearly demonstrated.

REFERENCES

1. B. Lewin, *Genes*, Wiley, New York, (1985).
2. D.H. Hamer and C.A. Thomas, The cleave of *Drosophila melanogaster* DNA by restriction endonucleases, *Chromosoma* **49**, 243–267, (1975).
3. K. Skriver, B. Tang and S. Bock, A bacteriophage λ cloning strategy employing 'completely' digested genomic DNA, (Preprint).
4. B. Tang and M.S. Waterman, The expected fraction of clonable genomic DNA, *Bull. Math. Biol.* **52**, 455–475, (1990).
5. M.S. Waterman, *Introduction to Computational Biology*, Chapman & Hall, (1995).
6. D.T. Bishop, J.A. Williamson and M.H. Skolnick, A model for restriction fragment length distributions, *Am. J. Hum. Genet.* **35**, 795–815, (1983).
7. M. McClelland and M. Nelson, Enhancement of the apparent cleavage specificities of restriction endonucleases: Application to megabase mapping of chromosomes, *Gene Amp. & Analy.* **5**, 257–282, (1987).
8. M.S. Waterman, Frequencies of restriction sites, *Nucl. Acids Res.* **24**, 8951–8956, (1983).
9. W. Kuhn, Über die Kinetik des abbaues hochmolekularer Ketten, *Ber Desch. Chem. Ges.* **63**, 1503–1508, (1930).
10. E.W. Montroll and R. Simha, Theory of depolymerization of long chain molecules, *J. Chem. Phys.* **8**, 721–727, (1940).
11. B. Seed, Theoretical study of the fraction of a long-chain DNA that can be incorporated in a recombinant DNA partial-digest library, *Biopolymers* **21**, 1793–1810, (1982).
12. B. Seed, C. Parker and N. Davidson, Representation of DNA sequences in recombinant DNA libraries prepared by restriction enzyme partial digestion, *Gene* **19**, 201–209, (1982).